International Journal of Database Theory and Application
Vol.8, No.2 (2015), pp. 235-244
http://dx.doi.org/10.14257/ijdta.2015.8.2.22

Extracting Entity Relationship Diagram (ERD) from English Sentences

Amani Abdel-Salam Al-Btoush

amanebtoush77@gmail.com

Abstract

Entity Relationship Diagram (ERD) is the first step in database design; it is an important step for the database designers, users, analyst, managers and software engineering. Since English is a universal language, this paper describes a methodology that extracts ERD from English sentences. This methodology is based on a predefined set of a heuristic rules that aims to extract the elements of the ERD, then these rules are mapped into a diagram. A diagram generator automatically converts the rules into the ERD according to the rules of generating. The proposed methodology is explained by examples to show how it can provide a mechanism for quickly and easily way in extracting the ERD.

Keywords: *Entity Relationship Diagram, entity, relationship, attribute*

1. Introduction

The Entity Relationship Diagram (ERD) shows that the real world consists of a collection of entities, the relationships between them, and the attributes that describe them. An entity is the object where we want to store data. A relationship defines the allowed connections between instances of entities [1]. Attribute is a characteristic common to all or most instances of a particular entity. Since the ER approach is easy to understand, a designer can focus on conceptual modeling of an organization, making decisions of what to use from entity sets, relationship sets and constraints [2].

The ER-Diagram tool provides a mechanism for quickly and easily modeling data structures required by a software system. The ERD tool provides all the usual features of a data modeling tool and additionally provides reverse engineering. Thus, the user can create a database system quickly on a number of different target platforms without the need to write any Data Definition Language (DDL) type code.

There are many essential concepts between ERD structure and English grammar structure [3] after analyzing the English sentences, so it is easy to make mapping between them. This paper describes a methodology, which can be able to extract the ERD from a description of the application domain given in English sentences. Using ER-extractor that extracts entities, relationships and attributes according to the heuristic rules it will defined, as well as by matching between the structures of both English and ERD structures. After that, ER-generator is depends on the predicate to convert the structure, which next pass the ER-descriptor to start to draw the ERD automatically depending on the rules.

This paper proposed to define a methodology that provide a help to the database designer to automatically extract the ERD from a given English sentences. ERD is the first step in database design, it is also a simple technique described in a graphical way to decide which database fields, relationships and tables will be the base of any database. ERD is a good communication tool between users and who use the system during the identification of the user information requirements process.

ISSN: 2005-4270 IJDTA

ER-Diagram Considered to be easy to understand for each of users, managers, analyst and database designers. ERD not only provides a modeling features but also it is the starting point of a safe and high quality database design with all of its well defined semantics [4].

The remainder of this paper is structured as follows: in the next Section a brief description of the methodology components. Section 3 described a real example of how the methodology can be applied into English sentences, Section 4 states the limitation. Related work is presented in Section 5, followed by conclusions and future work.

2. Methodology Overview

The proposed methodology of extracting the ERD depends on analyzing the English sentence. This aims to map each part of English sentence to what matched by the ERD component. English sentences used in this methodology must satisfy some rules to find the main components, each of these main component heuristically correspond to a concept in the ERD, each of these concept are then represented by a symbol and labeled by their associated name. The main structure of the proposed methodology is shown in Figure 1. This is being described according to the following subsections in details:

2.1. Sentences Checker

Sentence checker is the pre-process of the sentence tense, aims to check the English sentences that will be translated to the ERD. This step checks that sentence must satisfy the following Sentence Checker Rules (SCR):

SCR1: sentence must end by a full stop.

SCR2: check that each sentence must be in the past or present tense.

SCR3: the questions are not accepted.

The sentences that satisfy the checker rules will be accepted to move to the next step, other sentences will be ignored. For example if some one gives the sentence: How can I find it? This sentence will be ignored. Another sentence example: student pass exam, it is a past tense sentence but it does not end with a full stop, so also it will be ignored.

2. Sentence Component Finder

In this step English sentence is going throw steps to get the main component that are useful in ER-Diagram extraction. So the methodology dived this step into two sub-steps:

2.2.1. Sentence Part of Speech Finder: The need to know the part of speech the words belong to aims to use them correctly; English sentence has eight parts of speech, Nouns, Pronouns, Adjectives, Verbs, Adverbs, Conjunctions, Prepositions and Interjections. The main ER-Diagram components that this step aims to find from English sentence are a list of Nouns, Verbs, Adjectives and Adverbs, and save them as a table that contain the word and what part of speech it is, this table will be saved as a text file which then ready to use in the next step.

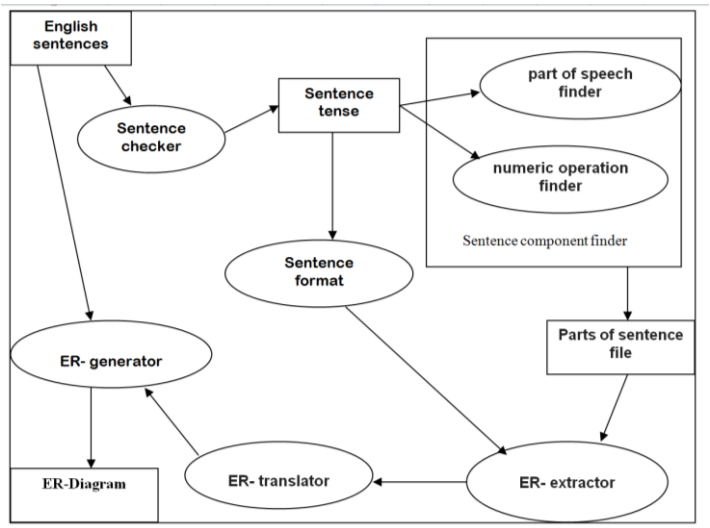


Figure 1. The Structure of Extracting ER-Diagram from English Text

2.2.2. Sentence Part of Numeric Operation Finder: Many of the numeric that are used in the real life take a format so this will easy dealing with them, as we know both Social security number and Phone number in each country take a format, so the proposed methodology can select the country as a first step of work, so it is easy to identify the format for both Social security number and Phone number that country use them. For example in Jordan If the format of the numeric is (9#########) name the attribute Social security number, and if the format of the numeric is (07#######) name the attribute Phone number. Add the word numeric and what name it takes to the table that establishes in step 2.2.1. The following sentence: "a person with a number 954237844 may own house and may belong to a political party.", will translate according to Table1, which will be saved as a text file.

Table 1. Example of the Equivalence between English Words and English Language Concepts

English sentence	English part of sentence	English language concepts
"A person with a number 954237844 may own house and may belong to a political party."	Person	Noun
	Own	Verb
	House	Noun
	Belongs to	Verb
	Political party	Noun
	Social security number	Numeric

2.3. ER- Extractor

ER- extractor aims to extract entities, relationships and attributes from the English sentence. Each English sentence will have its main (ERD) feature. Many rules were used to match between (ERD) features and English sentence part heuristically, based on works [4, 5, 6 and 7], Table 2 summarized these rules.

Table 2. The Correspondence between English Language Concepts and the ERD Features

English language concepts	ER-diagram feature
Common noun	Entity type
Transitive verb	Relationship type

Adjective	Attribute of an entity
Adverb	Attribute of Relationship
Numeric operation	Attribute of an entity
Proper noun	Entity type
Verb followed by a preposition	Relationship type
Noun followed by other noun belongs to the following set: { Name, Birth, Number, Type and Address}	both words are an Attributes

According to this step by applying the heuristic rules on Table 1, the output of ER-extractor summarized in Table 4.

English sentence	Part of English sentence	ER- extractor
" A person with a number 954237844 may own house and may belong to a political party."	Person	Entity type
	Own	Relationship
	House	Entity type
	Belongs to	Relationship
	Political party	Entity type
	Social security number	Attribute

Table 4. Example of the Correspondence between English Language Concepts and the ERD Features

Some English sentences take a form; in this form it will be easy to find the wanted component to make it easy to extract the ERD. So the base is the equivalent form, based on work [6], Table 3 summarized these format and how to deal with them.

English sentence form	The equivalent form
" The X of Y is Z", and Z is a proper noun.	X will be treated as a relationship between Y and Z, so both Y and Z considered to be an entities.
" There are … X in Y "	Y has X, so both X and Y are an attributes and "has" is the relation ship between them.
"The X of Y is Z", and Z is not a proper noun.	X will be treated as a relationship between Y and Z, so both Y and Z considered to be an entities.

Table 3. The Equivalent between English Sentence form and ERD For

2.3. ER- Translator

In this step the meaning of each English sentence must be represented in the First Order Logic (FOL), for this purpose we can use LISP as a programming language to translate English sentence into the FOL, then the ER-translator rules can be easily applied and because cardinalities are an important part of ERD. Applying cardinalities constraints on the output of ER-extractor is the main goal of step of ER-translator step, according to this there is need to use predicate.

The use of adjective "any" or "many" suggests a maximum cardinality. Also when the comparative adjective "more than" followed by a cardinality number, it indicate the degree of the cardinality between two entities. For example binary relationship

can have three possible cardinalities, (1:1) one-to-one, (1: N) one-to-many or (N: M) many-to-many. In ER- translator step, after translating English sentence into the correspondence FOL the following rules must be applied:

Rule 1: Every English sentence get a number ER (number), where number represent unique number for each sentence.

Rule 2: an entity written in this way E (E_num, E_name, S or P).
Where E: Entity, E_num: Number for identifying the Entity, E_name: String that represent entity's name, S for singular and P for plural.

Rule 3: the Relationship written in this way: R (R-num, R_name, E_s(S or P), E_d(S or P)). Where R: Relationship, R-num: Number for identifying the association, R_name: String for represent the Relationship Name, E_s: Source entity, E_d: Destination entity, S for singular and P for plural.

Rule 4: Attribute written in this way A (A_num, A_name, and A_src).
Where A_num: Number for identifying an attribute, A_name: String for represent an attribute name, A_src: source attribute.

According to this step, Table 4 firstly translated the sentence into the FOL, after that, the sentence represented according to ER-translator rules as shown in Figure 2.

English sentence: "A person with a number 954237844 may own house and may belong to a political party."

FOL representation: - ∀ x (person (x)) ➔ ∃ a (house (a), owns (x, a)).
- ∀ x(person (x)) ➔ ∃ a (political party (a, b), belong (x, a)).
- ∀ x(person (x)➔ security number (x)).

ER-representing: - ER (1): E (1, Person, S)
E (2, House, S)
E (3, Political party, S)
R (1, Own, E1(S), E2(S))
R (2, Belongs to, E2(S), E3(S))
A (1, Social security number, E1)

Figure 2. The FOL Translator and the Correspondence ER-representation of the English Sentence

2.5. ER- Generator

The ERD was introduced by many notations, the most accepted notation used to represent ERD, Chen's original notation. Chen's notation Symbolizes for the ERD attributes by oval symbols, relationship types are represented by a diamond between two entities and put the cardinalities, and the entities are symbolized by a box with entity name. According to the ER- translator, the symbol (P) represents in cardinality (N), and the symbol (S) represents in cardinality (1). If there are two cardinalities represented in the symbol (N) and related to the same relationship, one of them must be named (N) and the other named as (M).

3. Examples

To illustrate the steps of extracting ERD from English sentences, a list of examples were introduced. Table 5 gives some examples of finding the English language concept.

Table 5. Examples of Extracting English Language Concepts from English Sentences

Example	English sentence	Part of English sentence	English language concepts
1.	"A 30 year-old engineer works on a project with a project number 55612 for a percentage of his time."	Engineer	Noun
		Project	Noun
		Old	Adjective
		Project-number	Numeric
		Works-on	Verb
		For a percentage of time	Adverb
2.	"The average salary is $15,000."	The average salary	Numeric

Since the English language concept extracted from English sentence, it will be easily mapped into the correspondence ERD component using ER-extractor, according to the heuristic rules of part of speech finding, and the rules of numeric operation finding. Table 6 explains in examples how each part of English sentence from Table 5 is represented according the rules of representation.

Table 6. Examples of Mapping each Part of English Sentence to the Correspondence ERD Component

Example	English sentence	Part of English sentence	ER- extractor
1.	"A 30 year-old engineer works on a project with a project number 55612 for a percentage of his time."	Engineer	Entity type
		Project	Entity type
		Old	Attribute of an entity "Engineer"
		Project number	Attribute of an entity "Project"
		Works on	Relationship
		For a percentage of his time	Attribute of the Relationship "works on"
2.	"The average salary is $15,000."	The average salary	Attribute
3.	"Color of the car is green."	Car	Entity
		Color	Attribute
4.		Student	Entity

	"There are 25 students in the class."		
		Class	Entity
		Has	Relationship

As a previous step before drawing ERD, ER-translator step translate every English sentence as a predicate, so it translate the sentence into the First Order Logic (FOL). After that, the FOL sentence represented as ER-translator rules. So sentences in Table 6 represented as shown in Table 7.

Table 7. Examples of Predicate Translator for English Sentences

English sentence	Expression in FOL	ER-representation
1. A 30 year-old engineer works on a project with a project number 55612 for a percentage of his time.	- ∀ x(Engineer (x)) ➔ ∃ a (Project (a), works on (x, a)). - ∀x (Engineer(x)➔ old(x)). - ∀x (Engineer(x)➔ Project number (x)). - ∀x (Engineer(x)➔ percentage of time (x)).	ER(1): E(1, Engineer, S) E(2, Project, S) R(1, works on, E1(S),E2(S)) A (1, Old, E1). A (2, Project number, E2). A (3, percentage of time, R1).
2. The average salary is $15,000.	∃ a (The average salary (a)).	ER (2): A (The average salary).
3. Color of the car is green.	∀x (Car (x))➔ color (x)).	ER(3): E(1, Car, S) A (1, Color, Car).
4. There are 25 students in the class.	∀ x(Student (x)) ➔ ∃ a (Class (a), Has (x, a)).	ER(4):E(1, Student, P) E(2, Class, S) R (1, Has, E1 (P), E2(S)).

Discussion

The good ERD should satisfy some basic rules such as all relationships and attributes must be connected, the entity name should be unique, each entity at least should have one relation, it is impossible for a relationship to be connected directly to another relationship and for every entity there should be at least one attribute. The examples introduced here aims to show how the methodology applied on the English sentences, not to extract the good ERD that satisfy these basic rules.

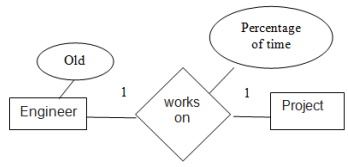
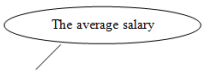
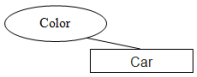
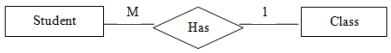
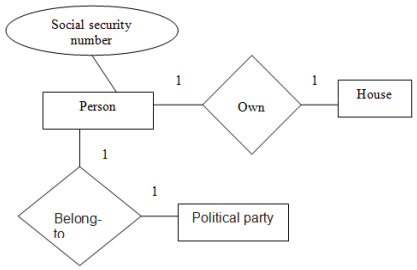
The introduced examples show the steps of extracting ERD from English sentences context, by applying a heuristics for language syntax rules. And representing the semantic of each English sentence depends on LISP as a programming language to translate every English sentence into the FOL. The steps

followed in the introduced methodology shows how it can provide a mechanism for quickly and easily extracting ERD automatically from English sentences.

Heuristic provide good solution to solve many difficult problems, not necessary the optimal solution. From this principle, the proposed methodology is based on the heuristic rules to extract the ERD component from English sentence.

As a last step in the proposed methodology the ER-generator depends on the generator rules and the translator rules to extract the ERD as shown in Figure 8

Table 8. Examples of English Sentence and the Correspondence ERD

English sentence	ER-Diagram
1. A 30 year-old engineer works on a project with a project number 55612 for a percentage of his time.	
2. The average salary is $15,000.	
3. The color of the car is green.	
4. There are 25 students in the class.	
5. A person may own house and may belong to a political party.	

4. Limitation

The heuristic rules are commonly used in extracting ERD from natural language processing, nevertheless, they can't rely support the interpretation. Also these heuristic rules were applied on a narrow range of English sentences.

The proposed methodology tries to give away that extract the ERD from a given English sentences, but the real word data model will contains a lot of entities, relationships and attributes. Another point considered to be a limitation that any English sentences concept can be represented in different ways, but Table 3 Restricted to take some cases of these sentences.

5. Related Work

Several approaches based on analyzing the relational database schema, Mfourga [8] describes a methodology that provides a framework for extracting an E-R schema from a set of form model schemas of the relational database. The framework supports recovering the problem of non-existence of the documentation. This methodology also explores the semantics contained in forms that serve as communication interface with databases. The information that is extracted from both form structure and instances where uses as a database reverse engineering input, the main strategy learning by example borrowed from the machine learning paradigm .

H. Alalfi, R. Cordy and R. Dean [9] use My-Sql implementation of the SQL data definition language (DDL). The aim of the tool is to depend on source transformation technology to fill the gap between application and data. This tool, called SQL2XMI, automatically changes an SQL schema into a UML-ER model using XML Meta Interchange (XMI) 2.1. By getting the data model to the UML world, this model builds UML class model to clarify the basic components of ER-Diagram components.

Dhabe and others [10] proposed a diagram to accommodate the functional dependency (FD) information as its integral part for complete automation of normalization. This system allows for a greater amount of automation, as well as for ARE diagram modification to be applied automatically to FD information. FD is diagrammatically represented using two types of connectors.

Shuyun Xu, Yu Li, and Shiyong Lu [2] used a tool that mixes between the relational database theory, OO technology, and XML to draw the ER-Diagram. The semantic drawing tool and verification process assure that ER-diagrams will be correct. To explain how this tool can be developed, this paper describes the architecture of ER-Draw and its application details.

None of the above works depends on the text to extract the ERD, The depend was on relational database schema to extract the ERD, or to enhance the ERD by applying the functional dependency on the ERD.

Other approaches depends on the natural languages analyzing , Shahbaz and other [11] propose an approach that aims to map the tagged words into the enhanced ERD by applying the Heuristically classifying into the natural language text.

Kouninef and Al-Johar [12] determine “Entities and Relationships” by starting with translation of Arabic text through the system’s components and getting the matching meaning in a first order logic form. After that, the tool extracts “Entities and Relationships” according to their rules, which describes the elements of the Conceptual Schema.

Naji [13] introduces a tool that builds Entity-Relationship Table, which splits the relationship from the subject item. Then these relations will be mapped to dictionary relationships to get real word presentation to the corresponding relationship words.

Batmaz and Hinde [14] propose a diagram tool for educational purposes, which aim to gain contextual information of all components in a conceptual database diagram from a given scenario. This tool enables user to input component types and names. Then the tool draws the diagram. No ERD will be drawn until the “Draw” button is pressed.

Manika Nanda [15] developed a framework that recognize the named entities from both English and Hindi language, the problem of saving the numbers and dates were solved by defining a different formats of date semi-automatic approach [16] described by analyzing natural language to obtain the entity relationship model depending on sentences logical form and applying the heuristics to identify the relationship.

This approach depends on analyzing natural languages to extract ERD, but many of the works either extracting the components of the ERD, or it does not extract ERD automatically while the proposed methodology aims to extract ERD automatically.

Conclusion and Future Work

Since the real world consists of a collection of entities and relationship between them with attributes that describes them, extracting ERD is an important step to understand the real word. ERD supplement data modeling, and also considered to be the base approach for database designers and software developers.

The proposed methodology solve the problem of extracting the ERD from a given English sentences using a set of proposed heuristic rules for entities, relationships and attributes, in a quick and easy way. Many examples were introduced to show how this methodology firstly finds the main component of English sentence. After that, a list of entities, relationships and attributes are extracted, then a diagram generator, which depends on the generating rules, converts the list of entities, relationships and attributes into the ERD automatically.

As the basis of automatically converting English sentences into ERD were introduced, the first future work we aims to setup a framework that extract the ERD from a given English text automatically, depending on the described methodology. The second future work is to improve the rules to be applied in a wider range.

References

[1] C. Helen, C. Purchase, R. Welland, M. McGill and L. Colpoys, "Comprehension of diagram syntax: an empirical study of Entity Relationship notations," International Journal of Human-Computer Studies, vol. 61, no. 2, **(2004)**, pp.187-203.
[2] S. Xu, Y. Li and S. Lu, " ER-Draw: An XML-based ER-diagram Drawing and Translation Tool," Computers and Their Applications, **(2003)**, pp.143-146.
[3] C. Peter, P. Shan, "English, Chinese and ER diagrams," Data & Knowledge Engineering, vol. 23, no. 1, **(1997)**, pp.5- 16.
[4] I. Song and Y. K. Froehlich, "Entity-relationship modeling," Potentials, IEEE, vol.13, no. 5, **(1994)**, pp.29-34.
[5] H. Sven and S. Link, "English sentence structures and EER modeling," Proceedings of the fourth Asia-Pacific conference on Conceptual modeling, Australian Computer Society, Inc, **(2007)**, vol. 67.
[6] C. Peter and P. Shan, "English sentence structure and entity-relationship diagrams," Information Sciences, vol. 29, no. 2, **(1983)**, pp.127-149.
[7] K. Christian, G. Flied and H. C. Mayr, "From Natural Language Requirements to a Conceptual Model," International Workshop on Design, Evaluation and Refinement of Intelligent Systems (DERIS2010), **(2010).**
[8] M. N, "Extracting entity-relationship schemas from relational databases: a form-driven approach," Reverse Engineering, 1997, Proceedings of the Fourth Working Conference on. IEEE, **(1997)**.
[9] A. H. Manar, J. R. Cordy, and T. R. Dean, "SQL2XMI: Reverse engineering of UML-ER diagrams from relational database schemas," Reverse Engineering, IEEE, **(2008).**
[10] M. Dhabe, S. Patwardhan, A. Asavari, A. Deshpande, M. L. Dhore, B. V. Barbadekar and H. K. Abhyankar, "Articulated entity relationship diagram for complete automation of relational database normalization", international journal of database management system, **(2010)** May, vol. 2, no. 2.
[11] M. Shahbaz, S. Ahsan, M. Shaheen, R. M. A. Nawab and S. A. Masood, "Automatic Generation of Extended ER Diagram Using Natural Language Processing," Journal of American Science, vol. 7, no. 8, **(2011).**
[12] B. Kouninef, B. Al-Johar, "Extracting Entities and Relationships from Arabic Text for Information System", Journal of Emerging Trends in Computing and Information Sciences, **(2011)** October, vol. 2, no. 11.
[13] A. M. Naji, "Proposal of Creating Entity-Relationship Table from English Sentences Groups", IJCCCE, **(2011)**, vol.1, no.2.
[14] O. Nazlia, J. R. P. Hanna and P. McKevitt, "Heuristic-based entity-relationship modeling through natural language processing," Artificial Intelligence and Cognitive Science Conference (AICS). Artificial Intelligence Association of Ireland (AIAI), **(2004)**.
[15] M. Nanda, "The Named Entity Recognizer Framework", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163, **(2014)** May, vol. 1, Issue 4.
[16] M. Farid, and S. Vadera, "Obtaining ER diagrams semi-automatically from natural language specifications." **(2004)**, pp. 638-642.

Author

Amani A AL-Btoush, was born in karak, Jordan 1984. she reseved BSc degree in computer science from Mutah University, Jordan in 2006. Currently, she is a master student in the Department of Computer science faculty of science at Mutah University.

